## fastQC指标

各项指标的颜色代表的含义：完全正常（绿），略有异常（橙） 或异常（红）

#### Basic Statistics （基础统计）

| Measure | Value |
| --- | --- |
| Filename | SRR10440857\_1\_val\_1.fq.gz |
| File type | Conventional base calls |
| Encoding | Sanger / Illumina 1.9 |
| Total Sequences | 33468481 |
| Total Bases | 4.9 Gbp |
| Sequences flagged as poor quality | 0 |
| Sequence length | 36-150 |
| %GC | 52 |

是一个基本统计模块包含所分析文件的基本统计信息。

* 文件名：被分析的文件的原始文件名
* 文件类型：说明文件calls类型
* 编码：表示在文件中找到的质量值是基于哪个 ASCII 编码
* 序列总数：处理的序列总数的计数
* 总碱基数：fastq文件中所包含的碱基总数
* 过滤序列：被标记为质量差的序列数目
* 序列长度：提供最短和最长序列的长度
* %GC：所有序列中所有碱基的总 GC含量

#### Per base sequence quality（每个碱基序列的质量）

![fig:](data:image/png;base64;base64,)

1.此图中的横轴是测序序列第1个碱基到第150个碱基；纵轴是质量得分（Q值），Q = -10\*log10（P error），即20表示错误率为1%，30表示错误率为0.1%。  
2.图中每1个boxplot，都是该位置的所有序列的测序质量的一个统计，上面的bar是90%分位数，下面的bar是10%分位数，箱子的中间的横线是50%分位数，箱子的上边是75%分位数，下边是25%分位数。  
3.图中蓝色的细线是各个位置的平均值的连线，一般要求此图中的所有位置的10%分位数大于20，也就是我们常说的Q20过滤。  
4.WARN (对应summary的⚠️) 出现的情况：任何碱基质量低于10，或者是任何中位数低于25；  
5.FAIL (对应summary的❌)：任何碱基质量低于5，或者是任何中位数低于20。

#### Per sequence quality scores（每个序列的质量分数）

![fig:](data:image/png;base64;base64,)

1.如何计算一条reads的质量值：假设该序列长度为150bp，那么这150个位置每个位置Q值的平均值就是该条reads的质量值。  
2.该图横轴是表示Q值；纵轴是每个值对应的reads数目。当测序结果主要集中在高分时，证明测序质量良好。  
3.当结果峰值小于27（错误率0.2%）时fastqc报“WRAN”，当峰值小于20（错误率1%）fastqc报“FAIL”。

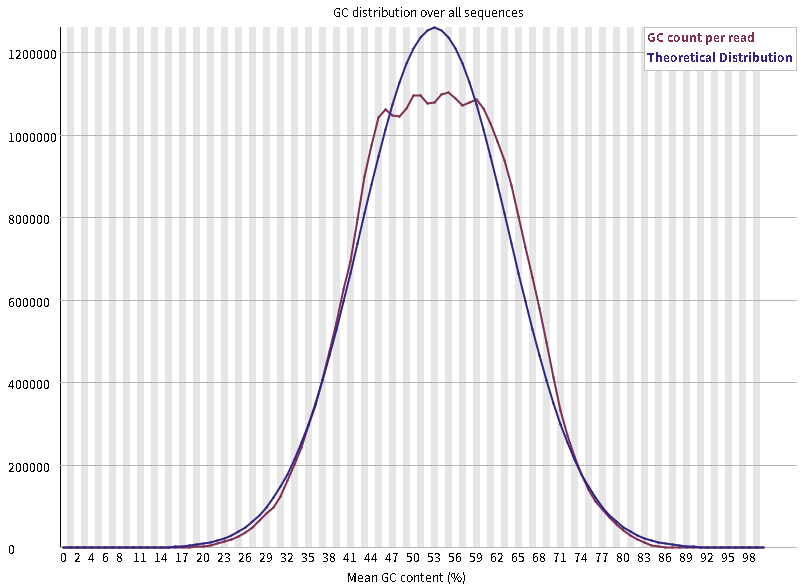
#### Per base sequence content（每个位置上的碱基的比例分布）

![fig:](data:image/png;base64;base64,)

对所有reads的每个位置，统计ATGC四种碱基的分布。这个图可以看出每条序列中各个位置的平均碱基比例，如果出现AT或GC分离的情况说明这个数据有问题，需要处理。由于AT配对，CG配对，假如测序过程是比较随机的话（随机意味着好），那么在每个位置上A和T比例应该差不多，C和G的比例也应该差不多。两者之间即使有偏差也不应该太大，最好平均在1%以内，除非有合理的原因，比如某些特定的捕获测序所致的偏差过高。

1.横轴是位置1 - 150 bp；纵轴是百分比。  
2.图中四条线代表A T C G在每个位置平均含量。理论上来说，A和T应该相等，G和C应该相等，但是一般测序的时候，刚开始测序仪状态不稳定，很可能出现严重分离的情况。像这种情况，即使测序的得分很高，通常需要修剪掉开始部分的序列信息，如该图需要cut前面10bp，可以使用TrimGalore等。  
3.正常情况下四种碱基的出现频率应该是接近的，而且没有位置差异。因此好的样本中四条线应该是平行且接近的。当部分位置碱基的比例出现bias时，即四条线在某些位置纷乱交织，往往提示有重复出现序列（overrepresented sequence ）的污染，当所有的碱基比例一致地表现出bias时，即四条线平行但分开，往往代表文库有bias（建库过程或本身特点），或者是测序中的系统误差。当任一位置的A/T比例与G/C比例相差超过10%，fastqc报“WRAN”，当任一位置的A/T比例与G/C比例相差超过20%，fastqc报“FAIL”。

#### Per sequence GC content（每条序列的GC含量）



GC含量指的是G和C这两中碱基占总碱基的比例。二代测序平台或多或少都存在一定的测序偏向性，我们可以通过查看这个值来协助判断测序过程是否足够随机。对于人类来说，基因组的GC含量一般在40%左右。因此，如果发现GC含量的图谱明显偏离这个值那么说明测序过程存在较高的序列偏向性，结果就是基因组中某些特定区域被反复测序的几率高于平均水平，除了覆盖度会有偏离之后，将会影响下游的变异检测和CNV分析。

1.横轴是0 - 100%； 纵轴是每条序列GC含量对应的数量。  
2.蓝色的线是程序根据经验分布给出的理论值，红色是真实值，两个应该比较接近才比较好。  
3.当红色的线出现双峰，基本肯定是混入了其他物种的DNA序列。  
4.偏离理论分布的reads超过15%时，报“WRAN”， 偏离理论分布的reads超过30%时，报“FAIL”。

#### Per base N content（序列中各个位点的N含量）

![fig:](data:image/png;base64;base64,)

1.序列中各个位点的N含量，越小越好。正常情况下N的比例很小，所以图上常常看到一条直线。当任一位置的N的比例超过5%，报“WRAN”， 当任一位置的N的比例超过20%，报“FAIL”。

2.N在测序数据中一般是不应该出现的，如果出现则意味着，测序的光学信号无法被清晰分辨，如果这种情况多的话，往往意味着测序系统或者测序试剂的错误。

#### Sequence Length Distribution（序列测序长度统计）

![fig:](data:image/png;base64;base64,)

1.每次测序仪测出来的长度在理论上应该是完全相等的，但是总会有一些偏差，数量比较少，不影响后续分析。  
2.当测序的长度不同时，如果很严重，则表明测序仪在此次测序过程中产生的数据不可信。  
3.当reads长度不一致时报“WRAN”，当有长度为0的read时报“FAIL”。经过修剪后的数据通常会报“WRAN”

#### Sequence Duplication Levels（统计序列完全一样的reads的频率）

![fig:](data:image/png;base64;base64,)

统计序列完全一样的reads的频率。测序深度越高，越容易产生一定程度的duplication，这是正常现象，但如果duplication的程度越高，就提示可能有bias的存在。

1.sequences duplication是指在测序前建库PCR过程中导致的一些序列扩增次数过多导致的。若重复较高则需要进行处理这些dup。  
2.横轴代表 reads 的重复次数 ( 1 表示 unique 的序列，2 表示有 2 条完全相同的 reads …)，大于 10 次重复后则按不同的重复次数合并显示。 纵坐标表示各重复次数下的 reads 数占总 reads 的百分比。  
3.曲线展示了所有 reads 的重复情况。  
4.需要注意的是，这个重复水平和建库方式有关。如果是 GBS 建库，由于存在 PCR 扩增的步骤，观察到的重复水平会比较高。  
5.fastqc中用fastq数据的前100,000条reads统计其在全部数据中的重复情况。当非unique reads占总数的比例大于20%时，报“WRAN”， 当非unique reads占总数的比例大于50%时，报“FAIL”。

#### Overrepresented sequences（重复出现序列）

Overrepresented sequences 显示同一条 read 出现次数超过总测序 reads 数的 0.1 % 的统计情况。如果[某种序列格外多，则证明有污染](https://www.jianshu.com/p/672e40fce5be)，污染来源有：实验中添加试剂（adapter或primer）；外源污染（人或细菌）。可以尝试使用过滤软件过滤，Trimmomatic、FASTX-Toolkit、SOAPnuke等。细菌污染则需要比对后再进行去除。

1.正常文库内序列的多样性水平很高，不会有同一条 read 大量出现的情况，这部分结果会把大量出现的 reads 列出来，并给出可能来源。  
2.在这部分分析中任何长度大于 75 bp 的序列都会被截成 50 bp。  
3.当发现超过总reads数0.1%的reads时报“WRAN”， 当发现超过总reads数1%的reads时报“FAIL”。

#### Adapter Content（接头含量）

![fig:](data:image/png;base64;base64,)

Adapter Content 显示 reads 中的接头含量，并显示可能的来源。

1.横轴为碱基位置，纵轴为含有接头序列的比例。  
2.如接头的含量随着碱基的位置增大而逐渐升高，则表示 reads 中含有接头，这部分接头会影响后续的分析，需要截掉 reads 中的接头序列或者将含有接头的 reads 完全删除。  
3.如果任何重复 read 超过总 reads 数的 5 % 则报 Warn， 超过总 reads 数的 10 % 则报 Fail。

### 总结

一般需要重点关注的主要是 Per base sequence quality、Per base sequence content和Adapter Content。

其中，如果Per base sequence quality太差的话，说明数据的质量远没有达到符合要求的Q30或着Q20的比例，这样测到的reads很多碱基是不可信的，对下游的分析结果影响比较大。

如果Per base sequence content的结果中出现很大异常的话(比如碱基的曲线出现明显的波动)，很可能提示原始下机数据中出现了很大比例的重复reads，这些重复的reads虽然本身的测序质量可能没有问题，可以使用TrimGalore进行修剪，有可能导致最终可用于分析的clean reads大大减少，需要引起注意。

如果Adapter Content参数曲线中，出现很大比例的adapter（接头）序列的话，可以使用TrimGalore、cutadapt和fastp，先根据接头序列去掉接头序列再进行分析。否则可能会影响后续的比对分析结果。